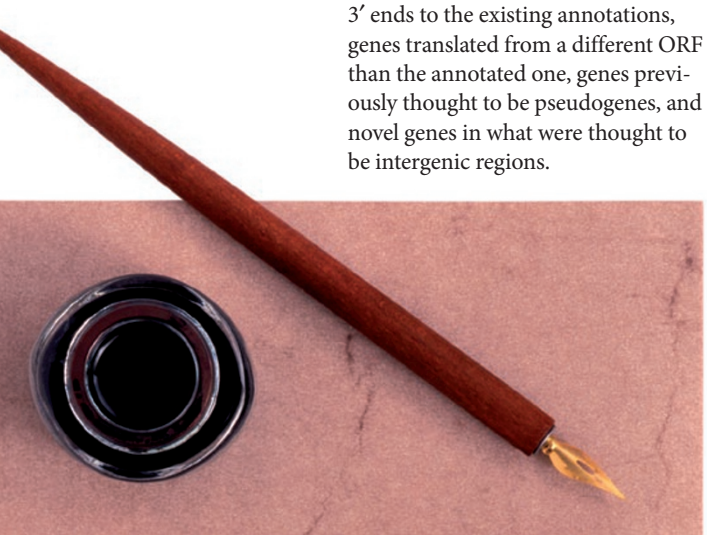
GENOMICS

Annotating with proteomes

The more genome sequences we have, the greater the need for accurate and efficient annotation. Proteomics offers a way of doing this that guarantees accuracy — if a protein is detected, then there must be a gene that encodes it. Two new proteomics studies have improved genome annotations by showing the existence of previously unidentified genes and by refining our knowledge of how known genes encode proteins.

Baerenfaller *et al.* used tandem mass spectrometry to characterize the proteomes of different *Arabidopsis thaliana* organs. The total set of 13,029 proteins corresponded to about half the genes in the annotated *A. thaliana* genome sequence. This protein set also provided 57 new or alternative annotations, among which are genes with different 5′ or 3′ ends to the existing annotations, genes translated from a different ORF than the annotated one, genes previously thought to be pseudogenes, and novel genes in what were thought to be intergenic regions.



The authors compared the proteomes from the different organs, and were particularly interested in the 571 proteins they identified in only one organ. They suggest that analysing the genes that encode these proteins might help identify the regulatory regions that direct organ-specific gene expression.

In the second study, rather than focusing on a single organism, Gupta *et al.* sought to combine the advantages of proteomic annotation with those of comparative genomics. They started by characterizing the proteomes of three related species of the bacterial genus *Shewanella*. For expressed proteins that did not correspond to annotated genes, or that suggested modifications of those annotations, they looked for sequence conservation across 10 other *Shewanella* genomes to confirm that their new annotation was accurate. In this way, they identified alternative start sites, rare protein modifications and programmed frameshifts (in which the genetic sequences do not predict protein sequences because structural effects during translation lead to a change in reading frame).

As the efficiency of high-throughput mass spectrometry improves, it is likely that proteomics will be used increasingly in genome annotation. As well as improving the accuracy of annotation, proteomics can provide information that other annotation methods are blind to, such as RNA editing and novel protein modifications.

Patrick Goymer

ORIGINAL RESEARCH PAPERS
Baerenfaller, K. *et al.* Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. *Science* 24 Apr 2008 (doi:10.1126/science.1157956) | Gupta, N. *et al.* Comparative proteogenomics: combining mass spectrometry and comparative genomics to analyze multiple genomes. *Genome Res.* 29 Apr 2008 (doi:10.1101/gr.074344.107)
FURTHER READING Reed, J. L., Famili, I., Thiele, I. & Palsson, B. O. Towards multidimensional genome annotation. *Nature Rev. Genet.* **7**, 130–141 (2006) | Brent, M. R. Steady progress and recent breakthroughs in the accuracy of automated genome annotation. *Nature Rev. Genet.* **9**, 62–73 (2008)